Yue Gao

□ +1 (608) 733-8789	isc.edu 🛛 🏠 pages.cs.wisc.edu/~gy	in ygao234	🞧 Lodour	🗣 Madison, WI
---------------------	-----------------------------------	------------	----------	---------------

RESEARCH INTERESTS

ML System Security, Adversarial Robustness, Attacks and Defenses, Security and Privacy. AI Security Vision Recognition, Natural Language Processing, Multi-Modality, Diffusion Models, LLMs. Machine Learning Multi-component ML Systems, Web-based Applications, Cryptography, Linux Kernel Memory. Cybersecurity

EDUCATION

University of Wisconsin-Madison

- Ph.D. Candidate in Computer Science (advised by Prof. Kassem Fawaz)
- Thesis: Challenges and Advances in Adaptive Response Strategies for Machine Learning Security
- Selected Courses: Introduction to Information Security, Applied Cryptography, Advanced Operating Systems.

Shanghai University

B.S. in Computer Science and Technology (GPA 3.99/4.00, Ranked 1/292)

- Thesis: A Deep Neural Network based Image Compression Method
- Selected Courses: Operating Systems, Computer Network, Assembly Language, Software Engineering.

WORK EXPERIENCE

C	
Advised by Prof. Kassem	n Fawaz

- Researched the security of **multi-component ML Systems** deployed in real-world environments.
- Systematized the security analysis of ML-based and web-based systems in **black-box** settings.

ML System Security Research Intern @ Microsoft Research

Research Assistant @ University of Wisconsin-Madison

Mentored by Dr. Jay Stokes and Dr. Emre Kiciman

- Proposed a research project on defenses against imperceptible textual **backdoor attacks on language models**.
- Discovered blind spots in state-of-the-art attacks and defenses, and published stronger defenses at MILCOM.
- Successfully reduced the attack success rate from 100% to 12%, even at a challenging poisoning rate of 10%.

ML Research and Development Intern @ TuCodec (Startup)

- Mentored by Dr. Chunlei Cai
- *ML* Research
- Secured 1st place as a **primary contributor** in the CVPR 2018 Challenge on Learned Image Compression.
- Improved the average runtime of DNN-based image compression algorithms from 1 min to 4 secs per 4K-res image.
- ML Engineering & Security
- Independently developed ML-based desktop apps on Ubuntu, MacOS, and Windows using TensorFlow and C++.
- Independently developed ML-based image compression systems using TensorFlow, Python, Docker, and Kubernetes.
- Protected ML systems from security intrusion and model stealing with security measures and anomaly analysis.

PUBLICATIONS

Conference

- [1] On the Limitations of Stochastic Pre-processing Defenses Yue Gao, Ilia Shumailov, Kassem Fawaz, and Nicolas Papernot Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS), 2022
- [2] Rethinking Image-Scaling Attacks: The Interplay Between Vulnerabilities in Machine Learning Systems Yue Gao, Ilia Shumailov, and Kassem Fawaz Proceedings of the 39th International Conference on Machine Learning (ICML), 2022 Oral Presentation (Top 2%)

Shanghai, China

Redmond, WA

Jun 2021 – Sep 2021

Shanghai, China Jan 2018 – Jul 2018

Sep 2018 – Mar 2024

Madison, WI

Sep 2014 – Jul 2018

Madison, WI Nov 2018 – Present

Yunang Chen*, Yue Gao*, Nick Ceccio, Rahul Chatterjee, Kassem Fawaz, and Earlence Fernandes 31st USENIX Security Symposium (USENIX Security), 2022
Bug Bounty (\$1500)
[4] I Know Your Triggers: Defending Against Textual Backdoor Attacks With Benign Backdoor Augmentation

[3] Experimental Security Analysis of the App Model in Business Collaboration Platforms

[4] **I Know Your Triggers: Defending Against Textual Backdoor** Attacks with Benigh Backdoor Augmentation Yue Gao, Jack W. Stokes, Manoj Prasad, Andrew Marshall, Kassem Fawaz, and Emre Kiciman IEEE Military Communications Conference (MILCOM), 2022

Workshop

[1] Variational Autoencoder for Low Bit-rate Image Compression
 Lei Zhou*, Chunlei Cai*, Yue Gao, Sanbao Su, and Junmin Wu
 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2018

Winner of the first Challenge on Learned Image Compression

Preprints

[1] SEA: Shareable and Explainable Attribution for Query-based Black-box Attacks

Yue Gao, Ilia Shumailov, and Kassem Fawaz *arXiv*, 2023

- [2] Human-Producible Adversarial Examples David Khachaturov, Yue Gao, Ilia Shumailov, Robert Mullins, Ross Anderson, and Kassem Fawaz arXiv, 2023
- [3] Analyzing Accuracy Loss in Randomized Smoothing Defenses Yue Gao*, Harrison Rosenberg*, Kassem Fawaz, Somesh Jha, and Justin Hsu arXiv, 2020

SELECTED PROJECTS

Real-world ML Systems Security: Threat Modeling, Defending, and Characterization Sep 2020 – Jun 2023

- Investigated the security of **multi-component ML systems** exposed to diverse security threats, e.g., dependencies.
- Revealed threats amplified by 9x and broke state-of-the-art defenses by jointly exploiting multiple vulnerabilities.
- Formally proved the non-robustness of randomization-based defenses beyond demonstrating empirical attacks.
- Characterized the attack's progression for **forensic purposes** and **human-explainable intelligence sharing**.

GARD: Guaranteeing AI Robustness Against Deception

Jun 2019 – Feb 2024

- Leadership
- Led a 9-member cross-university team to 1st and 2nd place in grant competitions hosted by DARPA.
- Developed initial code bases and coordinated **team members with varying technical backgrounds**.
- Onboard and mentored new team members to maintain adversarial mindsets with designing defenses.
- Individual Contribution
- Performed red teaming and broke over 10 internal defenses proposed by team members prior to submission.
- Proposed CLIP-like and diffusion-based methods to **enforce robust features** across RGB and Depth modalities.
- Successfully reduced the disappearance rate from 62% to 9% even **under the red-team evaluation from MITRE**.
- Contributed plug-and-play modules to the official upstream evaluation team and received acknowledgment.

Security Analysis of Online Business Collaboration Platforms

Mar 2021 – Dec 2021

- Automated the analysis of security principle violations for 3K+ third-party apps in Slack and Microsoft Teams.
- Reverse engineered OAuth designs to bypass access control, and received bug bounty for medium severity.
- Demonstrated POC attacks of eavesdropping on private chats, spoofing video calls, and unauthorized code merging.

Selected Honors & Awards

Slack Bug Bounty: Medium Severity, \$1500
Top 10% Reviewers Award: NeurIPS
CVPR Competition Winner: Challenge on Learned Image Compression
National Scholarship: China

Top 100 Elite Collegiate Award: China Computer Federation Scholarship for Exceptional Leadership: Shanghai University City Scholarship: Shanghai Outstanding Student Award: Shanghai University Outstanding Volunteer Award: ACM ICPC Asia Regional Contest Scholarship for Exceptional Innovation: Shanghai University Scholarship for Exceptional Academic Achievements: Shanghai University Bronze Prize for Programming Contest: ACM ICPC Asia East-Continent Final Contest Bronze Prize for Programming Contest: ACM ICPC Asia Shanghai Regional Contest		
PROFESSIONAL ACTIVITIES		
Reviewer: NeurIPS, ICML, and ICLR External Reviewer: IEEE Symposium on Security and Privacy External Reviewer: USENIX Security Symposium External Reviewer: ACM Conference on Computer and Communications Security Team Leader: Collegiate ICPC Team at Shanghai University	2022 - 2024 2021 - 2023 2021 - 2022 2019 2016 - 2017	
TALKS		
 Forensics and Intelligence Sharing for ML Security DARPA GARD PI Meeting, IBM Research The Vulnerabilities of Preprocessing in Adversarial Machine Learning ML Pad Team. Coords 	Oct 2023 Oct 2023	
 3. The Vulnerabilities of Preprocessing in Adversarial Machine Learning TrustML Young Scientist Seminar, RIKEN AIP 	Apr 2023	
4. On the Limitations of Stochastic Pre-processing Defenses DARPA GARD PI Meeting, University of Southern California (virtual)	Oct 2022	
5. The Interplay Between Vulnerabilities in Machine Learning Systems DARPA GARD PI Meeting, University of Michigan	Sep 2022	
6. Experimental Security Analysis of the App Model in Business Collaboration Platforms USENIX Security 2022	Aug 2022	
7. The Interplay Between Vulnerabilities in Machine Learning Systems ICML 2022	Jun 2022	
TEACHING AND MENTORING		

Project Mentor: DARPA GARD Project, University of Wisconsin-Madison	Fall 2023
Teaching Assistant: CS 368 (C++ for Java Programmers), University of Wisconsin–Madison	Fall 2018
Guest Lecturer: Advanced Algorithms & Data Structures, Shanghai University	2015 – 2017
Problem Designer: Undergraduate Programming Contests, Shanghai University	2015 – 2017
Student Mentor: Undergraduate Computer Science Coursework, Shanghai University	2015 – 2017

TECHNICAL SKILLS

Python	Expertise in ML and Security Research ^(2018 – present) and Backend Development ^(2016 – 2018) .
PyTorch	Expertise in ML Research ^(2019 - present) and Distributed Training ^(2020 - present) .
Docker	Expertise in ML Research (2018 - present) with familiarity in Kubernetes Cluster (2017 - present).
C / C++	Proficient in Security Research ^(2018 - 2019) , Software Development ^(2017 - 2019) , and ICPC ^(2014 - 2018) .
Security	Familiarity in IDA Pro, OllyDbg, Burp Suite, and nmap for CTF ^(2015 – 2017) .
TensorFlow	Familiarity in ML Research (2017 – 2020) and Service Deployment (2018).

ARTICLES AND MEDIA COVERAGE

CleverHans. Can stochastic pre-processing defenses protect your models?2022USENIX login. Experimental Security Analysis of the App Model in Business Collaboration Platforms2022Wired. Slack's and Teams' Lax App Security Raises Alarms2022